

Analysing differences among animal songs quantitatively by means of the Levenshtein distance measure

Jakob Tougaard^{1,2,3)} & Nina Eriksen^{1,4)}

(¹ Institute of Biology, University of Southern Denmark, Odense; ² National Environmental Research Institute, Roskilde, Denmark; ⁴ Research Laboratory for Stereology and Neuroscience, Bispebjerg University Hospital, Copenhagen, Denmark)

(Accepted: 1 December 2005)

Summary

The Levenshtein or string edit distance is an objective measure of the difference between two strings of elements. Levenshtein distance analysis has previously been applied to humpback whale songs, where it provided a quantitative measure of song change from year to year. This analysis is extended and a first approach to a robust statistical test is developed. The statistical test addresses the central issue whether two groups of songs (either from different individuals, different groups or different years) belong to the same population of songs or are significantly different. This is accomplished through derivation of the Kohonen median song sequence, which has the smallest possible summed Levenshtein distance to all songs of the group. By a simple *t*-test or nonparametric equivalent it is tested whether the median distance to the Kohonen median song sequence of a second group is significantly larger, which indicates that the groups are different. The test is expanded to handle multiple comparisons among several groups of songs.

Keywords: Levenshtein distance, string edit distance, song, Humpback whales, *Megaptera novaeangliae*, birdsong.

³⁾ Corresponding author's address: National Environmental Research Institute, Dept. of Arctic Environment, Frederiksborgvej 399, P.O. Box 358, DK-4000 Roskilde, Denmark; e-mail address: jat@dmu.dk

Introduction

A common problem in studies of song communication, regardless of which animal is producing the song, is to quantify differences and similarities between individual songs. Quantitative measures are central whenever one seeks to describe phenomena such as individual variation, geographical variation, and cultural and genetic evolution of the songs. More specifically, this comes in the form of questions such as whether songs of a particular individual or group of individuals are significantly different from another individual or group. Quantitative answers to such questions require quantitative measures of song difference.

Songs can be accurately described and compared by their acoustic waveforms and frequency spectra, yet such a description tends to overemphasise fine scale differences in time and frequency. This may be desirable for some types of song, especially in insects (e.g., Helversen & Helversen, 1998; Schul, 1998), but for much communication in vertebrates it is necessary to focus on similarities between songs on a larger scale, such as frequency modulations and relative duration and repetition rate of individual elements in the analysis.

A well-established approach to analysing songs based on large-scale similarities is to separate each song into individual units, phrases or syllables (terminology depends on animal group), each with certain distinct and stereotypical features (e.g., birds: Thorpe, 1954; Jenkins, 1977; humpback whales: Payne & McVay, 1971; Chabot, 1988; and most recently also mice: Holy & Guo, 2005). This is most commonly done by a more or less subjective evaluation of spectrograms of the songs. The use of subjective or quasi-objective methods in the initial classification obviously poses several problems for the subsequent analysis, but it is beyond the scope of this work to enter a discussion of this. In the following, we will simply take as starting point that a reasonably robust and unbiased analysis of the songs has resulted in characterisation of a number of individual, identified units. Each song can thus be represented as a sequence (string) of letters or symbols, each representing one unique unit.

In the analysis of humpback whale song Helweg et al. (1998), introduced and measured Levenshtein distances (see below) between song sequences in order to quantify similarities and differences between songs from different years. This was followed up in a recent study of humpback songs from Tonga

(Eriksen et al., 2005). The aim of the following is to elaborate on this work, and to describe in details the statistical method used for analysing differences between groups of songs based on Levenshtein distances.

Levenshtein distance

The Levenshtein distance, also known as the string edit distance, is a general measure of similarity between two text strings, and is thus readily applied to songs that can be represented as sequences of individual units or elements, each assigned a unique character. The Levenshtein distance between two strings of characters, here following the definitions of Kohonen (1985; 1988), express the minimum number of characters to be deleted, inserted or substituted in order to convert the first string into the second (or vice versa, as the process is reversible). Levenshtein distance comes in two general forms, the simple (unweighted) Levenshtein distance and the weighted Levenshtein distance. The simple Levenshtein distance $LD(a, b)$ is nothing more than a count of the minimum number of operations (insertions, deletions and substitutions) needed to convert string a into string b :

$$LD(a, b) = \min(i + d + s)$$

For example, the difference between the strings LEVENSHTein and LICHTENSTEIN is 5 (two insertions: I and C; two substitutions: E with H and V with T; and one deletion: H). In the weighted Levenshtein distance, each operation is multiplied by an individual weight before the sum to be minimised is calculated. In the most general form, three matrices represent the weights, and each contains individual weights for all combinations of characters. The introduction of individual weights allows for certain operations (e.g., the insertion of certain characters) to be favoured over others, and others to be suppressed or even prohibited. The individual weights can either be assigned based on some independent *a priori* knowledge on mechanisms underlying natural song changes or variation, or they can be generated in a bootstrap fashion from the dataset itself. In the latter case a very large dataset is required, and the system must be assumed to be ergodic (i.e., change assumed to be gradual and with constant rate). For well-studied song systems, as some types of bird song, the weighted Levenshtein distance may hold some promise, but for systems such as whale song where datasets are small and *a priori* knowledge on mechanisms even smaller, the simple Levenshtein

distance seems more appropriate. Only the simple Levenshtein distance is applied in the following. For further discussion of Levenshtein distances and practical algorithms for calculation of these see Kohonen (1988) and Sankoff & Kruskal (1983).

Methods

Testing for differences between sets – an example

For some analyses a simple comparison of Levenshtein distance between pairs of strings (songs) may suffice for conclusions to be drawn, but in most cases there is a need for proper hypothesis formulation and subsequent statistical tests of these hypotheses. Often one will ask questions like “are the songs in group A significantly different from the songs in group B?”. The groups of songs can be from different individuals or divided based on geography, time or any other parameter of interest. In any case, we have one group of songs, which we want to compare to one or more different groups of songs.

In the Levenshtein distance we found an objective measure of distance between two strings, and thus individual pairs of songs. A central obstacle for testing statements as the above however, is the realisation that Levenshtein distances are not arithmetic quantities. In other words, they do not add up in a proper way. If the Levenshtein distance between string a and string b is n and the Levenshtein distance between b and c is m , then it does not follow that the distance between a and c is always $n + m$. It gets even worse, because it is not possible to order (rank) a set of sequences in a sensible manner according to their pairwise Levenshtein distances. Take four sequences: ABC, ADC, AEC and AFC, as an example. There are six possible pairs and the distance between sequences in all pairs is 1 (one substitution of the second character). It is thus not possible to order the four sequences so that Levenshtein distances between neighbours are smaller than between next-neighbours. There simply is no natural order in which the four sequences can be arranged, according to their pairwise Levenshtein distances.

From this follows that standard statistical methods cannot be applied to the Levenshtein distances directly. It can be useful to think of Levenshtein distances as straight line distances between pairs of sequences, arranged

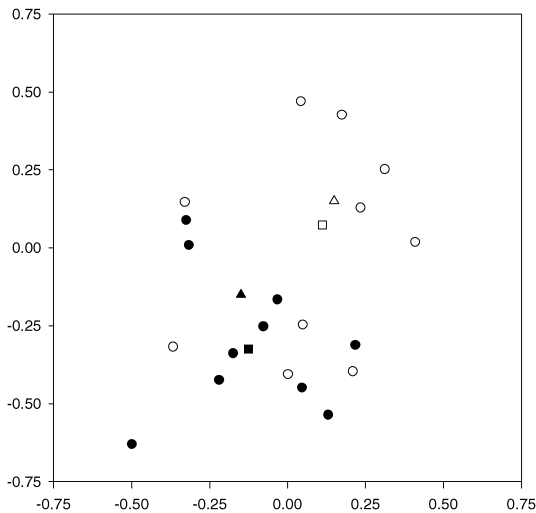


Figure 1. Two sets of randomly selected points (solid and open circles), with Kohonen medians of each set indicated by squares. The two sets were randomly drawn from two different bivariate Gaussian distributions with means indicated by triangles. In case of the bivariate Gaussian distribution the Kohonen median equals the mean.

not along a single dimension (line), but scattered as points in an $(m - 1)$ -dimensional space, where m is the total number of sequences involved.

To illustrate this and outline a solution, we turn to a simpler, yet analogous two-dimensional situation, distribution of groups of points in the plane. Figure 1 shows two sets of 10 randomly selected points. A reasonable question to ask is whether the two groups are random subsamples of the same population of points in the plane, or if the two groups are samples from different populations.

The following approach to testing this question relies on the set median. The set median M of the set A (following the definition by Kohonen, 1985) is the point (x_m, y_m) which has the minimum summed distance to all points of the set:

$$M_A : \min \sum_A \sqrt{(x_i - x_m)^2 + (y_i - y_m)^2}$$

This definition of the median differs somewhat from the median of conventional statistics and is referred to as the Kohonen median in the following.

The null-hypothesis H_0 of the test is: *Sets A and B are randomly drawn subsets of some larger set H.* If H_0 is true, this implies that both M_A and

M_B are estimates of the true Kohonen median M_H . Thus we could test H_0 by testing whether the Kohonen medians of the two sets, M_A and M_B are significantly different. This is where we run into problems, because the elements of the two sets A and B cannot be ranked and we thus cannot speak in a meaningful way of one Kohonen median being larger than the other. Instead we exploit the fact that the summed distance from the Kohonen median to all elements of a set is the minimal possible of all points in space (by definition). Any other point that is an equally good estimate of the true Kohonen median should have a summed distance to the set elements not significantly larger than the summed distance to the median estimated from the set. Thus, for set A we test whether the average distance to set A 's own Kohonen median (M_A) is significantly smaller than the average distance to set B 's Kohonen median (M_B). This is a straightforward test, as we can easily calculate mean distances to M_A and M_B and compare the two means in a t -test or equivalent non-parametric test, if required. The test is one-sided, as the mean distance to M_B by definition cannot be smaller than the mean distance to M_A .

It is tempting to conduct the test as a paired test, as we have two measures for each element in the set: distance to M_A and distance to M_B . This would be incorrect however, as the central assumption of a paired test is a correlation between the two elements of each pair. Because we are dealing with points in the plane and not points on a line, it does not follow that two points from set A with the same distance to M_A also has the same distance to M_B . Thus the test is unpaired.

The test performed is a test of whether M_A and M_B are equally good estimators of the true Kohonen median for set A . In order to complete the test, we need to repeat the process for set B , i.e. test whether the mean distance from the elements of B to M_A is significantly larger than the mean distance to M_B .

Most often the two differences (M_B against set A and M_A against set B) will either both be significant or both non-significant and it is immediately evident whether H_0 should be rejected or not. Inevitably however, some cases will turn up with one difference being significant and the other not. This is not a paradox, it simply informs us that we have committed either a type I error (the mean distance from one of the Kohonen medians to elements of one set is significantly different from the mean distance to the elements of the other set, although H_0 is true) or a type II error (the mean distance from

Table 1. Results of testing the null-hypothesis that set A and set B from Figure 1 are random sub-samples of the same set H. In reality, they were sampled from two different bivariate Gaussian distributions.

	Set A	Set B
True mean of bivariate Gaussian distribution	(−0.15, −0.15)	(0.15, 0.15)
Estimate of Kohonen median	(−0.03, −0.23)	(0.21, 0.17)
Mean distance to own Kohonen median (SD)	0.38 (0.18)	0.34 (0.12)
Mean distance to other Kohonen median (SD)	0.64 (0.23)	0.59 (0.25)
P_t -test, one-tailed, unpaired	0.011	0.006

the other Kohonen median to the elements of the two sets not significantly different, although H_0 is false). There is no direct solution to this dilemma, but an approach to overcome it is to assess the power of the two tests (if possible) and make the final decision based on the test with the highest power.

Table 1 shows the results of a test on the data from Figure 1. As both tests are significant on the 5% level, we reject the null-hypothesis and conclude that the two sets are sampled from different populations, as indeed they were.

Comparison of Humpback whale songs

Faced with two sets of strings, with characters of the strings representing units of animal songs, such as humpback whale song, we can proceed in a similar way as in the two-dimensional example above. In the Levenshtein distance we have an objective measure of distance between pairs of songs and we can define a Kohonen median song sequence for a set of songs in the same way as above. The Kohonen median song sequence for a given set is thus the sequence of units with the minimal summed Levenshtein distance to all the song sequences of the particular set. This sequence can be identical to one of the actual songs of the set, but this is not required (analogous to the Kohonen median points in the example above not being identical to any of the points of the two sets). We can now calculate the distance from each song sequence of a set to the set's own Kohonen median as well as the distance to the Kohonen median of the other set, and conduct the double test as shown above for the point example.

The approach outlined above can also be applied to more than two sets of strings. In that case sets are compared pairwise as above, with the addition

that probabilities should be properly corrected for the multiple comparisons performed (e.g., Bonferroni/Dunn-Sidak correction, Sokal & Rohlf, 1995), but see also discussion below).

As an example, the analysis of humpback whale song development done by Eriksen et al. (2005) is expanded in details below. Eriksen et al. (2005) should be consulted for details on methods of recording and analysis of songs. Important in this context is that a number of songs were recorded in the waters off Tonga in the years between 1993 and 1998. Each song was divided up into sequences of units (phrases), based on time-frequency contours and each sequence grouped into one of nine classes (themes), assigned A to H (with a variant of theme A named A₂). In the analysis presented in Eriksen et al. (2005) songs from 1991 were also included, but as there was no overlap in sequence groups (themes) between 1991 and the following years, these songs have been excluded in this context for reasons of increased clarity.

Results and discussion

For each of the six years a Kohonen median song sequence was constructed and the Levenshtein distances from all songs of that particular year to the Kohonen median song were compared to Levenshtein distances from all other songs to the Kohonen median, done year by year. The three sequences recorded in 1993 can serve as an example. The sequences were ACDEFGFGF, ACDEFG and ACDEFCFGFG. To find the sequence of the three closest to the Kohonen median we calculate the Levenshtein distance between the three combinations of pairs and for each find the sum of differences to the other two sequences (Table 2). The sequence ACDEFGFGF has the smallest sum of differences and is thus the first candidate for the Kohonen median sequence. To test for the existence of a hypothetical fourth song with a smaller summed distance to the three sequences of 1993, a systematic search through all possible insertions, deletions and substitutions to the sequence ACDEFGFGF was undertaken by means of custom written software, based on the algorithms devised by Kohonen (1988). It turns out that no such song exists and the song ACDEFGFGF thus qualify as the Kohonen median for the three 1993 songs.

Kohonen median song sequences were found in the same way for the remaining years and these were systematically compared to the recorded

Table 2. Levenshtein distances between song sequences recorded in 1993. Middle sequence has the smallest summed distance to the other songs and is identical to the Kohonen median (see text for explanation).

Song	Comparison song			Sum of Levenshtein distances
	ACDEFG	ACDEFGFGF	ACDEFCFGFG	
ACDEFG	0	3 insertions (FGF)	2 + 2 insertions (CF + FG)	7
ACDEFGFGF	3 deletions (FGF)	0	2 insertions (C + G)	5
ACDEFCFGFG	2 + 2 deletions (CF + FG)	2 deletions (C + G)	0	6

songs of the other years. Comparison of year 1993 and 1996 can serve as an example (Table 3). Levenshtein distances between the three songs from 1993 and the Kohonen medians of 1993 and 1996 are tabulated to the left and similar distances for a subsample of the 36 songs from 1996 to the right. Results of comparisons across all combinations of songs and Kohonen medians are shown in Figure 2.

In all cases the songs from the same year as the Kohonen median used for comparison has the lowest 50% percentile Levenshtein distance to the Kohonen median, which follows from the definition. The Levenshtein distance to Kohonen median increases for increased separation in time between songs and Kohonen medians. This is illustrated in Figure 3, which shows a clear separation of songs with time, consistent with the conclusions of Eriksen et al. (2005) that a gradual evolution of songs took place over the years.

In order to test the robustness of this conclusion, a series of pairwise test were performed, in line with the example above. As the distributions of Levenshtein distances are highly asymmetric and with unequal variance (evidenced by the error bars in Figure 2), a nonparametric test, the Mann-Whitney/Wilcoxon test was selected. One example of comparisons is shown in Table 3, where songs from 1993 and 1996 were compared. In both cases the average distance to the Kohonen median of the same year as the songs was considerably smaller than average distance to the Kohonen median of the comparison year. This difference was highly significant in case of the 1996 songs but due to the low number of samples in 1993 ($N = 3$) the p -value for comparing 1993 songs to the 1996 Kohonen median is considerably larger and becomes insignificant when corrected for multiple com-

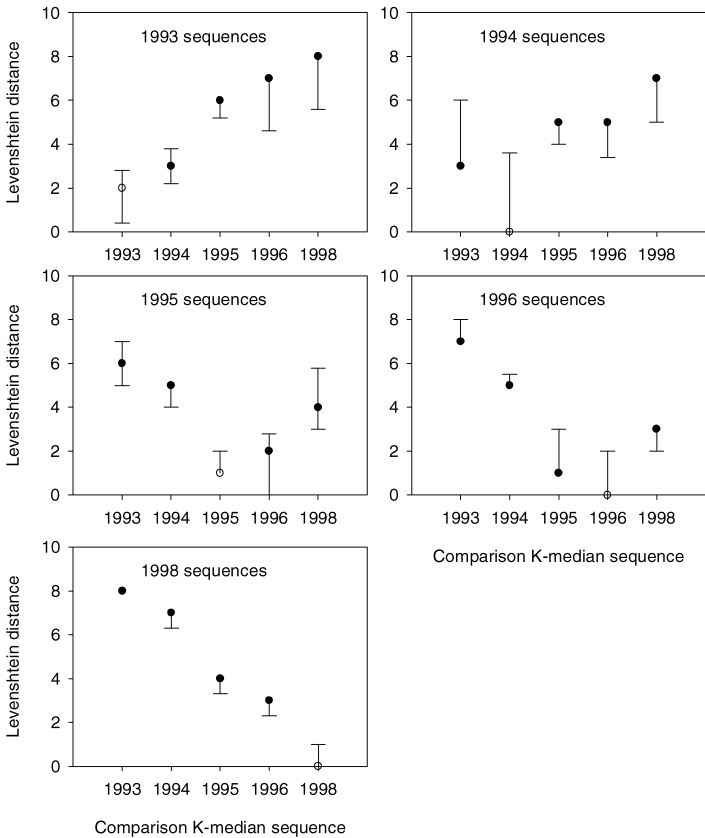


Figure 2. Levenshtein distances, expressed as 50% percentiles, between all songs of a particular year to the Kohonen median sequence of that year (open circle) and Kohonen median sequences of all other years (solid circles). Error bars represents 10% and 90% percentiles.

parisons (Table 4). As the power of the comparison between 1993 songs and 1996 Kohonen median is very low due to the low sample size of three, weight should be put on the test between 1996 songs ($N = 36$) and 1993 Kohonen median which was highly significant, also following correction for multiple comparisons (Table 4) and we conclude that songs in 1993 and 1996 are different. Results of all pair-wise comparisons are shown in Table 4. In general the results support the hypothesis that songs from different years represents different groups of songs. There are two exceptions. Songs from 1993 were not significantly separated from the Kohonen median songs of any of the other years, whereas the opposite was the case. As stated above,

Table 3. Comparison of songs from 1993 and 1996. Left columns show comparison of Levenshtein distances from 1993 songs to Kohonen median songs from both years. Right columns show comparison of Levenshtein distances from 1996 songs to Kohonen median songs from both years. Bottom rows contain means and standard deviations and results of the Mann-Whitney/Wilcoxon nonparametric test between distances to own Kohonen median and Kohonen median of the other year. *p*-values are not corrected for multiple comparisons.

1993 songs vs 1996 median				1996 songs vs 1993 median			
No. Sequence		Distance	Distance	No. Sequence		Distance to	Distance to
		to 1993	to 1996			1996	1993
		K-median	K-median			K-median	K-median
1	ACDEFGFGF	0	7	1	AA ₂ BCFH	0	7
2	ACDEFG	3	4	2	AA ₂ BCFGFH	2	5
3	ACDEFCFGFG	2	7	3	AA ₂ BCFGFH	2	5
				4	AA ₂ BC	2	8
				5	AA ₂ BC	2	8
				6	AA ₂ BCFH	0	7
				⋮			
				36	A ₂ B	4	9
Mean distance		1.7	6.0	Mean distance		0.8	7.2
SD		1.5	1.7	SD		1.1	0.7
<i>p</i> -value (uncorrected)		0.046 (<i>N</i> = 3)		<i>p</i> -value (uncorrected)		<0.0001 (<i>N</i> = 36)	

this is most likely due to the low sample size in 1993 and hence low power of the tests. Weight was not be put on these tests and we conclude from the reverse tests that songs from 1993 were different from songs of all other years. In a similar way the songs of 1995 were not significantly separated from the Kohonen median song of 1996, whereas the opposite was the case. The 1995-1996 comparison is more difficult than the case of the 1993 songs, as sample sizes are reasonably large for both 1995 and 1996 songs. There is no formal way to calculate power of a Mann-Whitney/Wilcoxon test (the probability of a type II error) and the question whether 1995 and 1996 songs are significantly different is thus left unresolved. See however Eriksen et al. (2005) for an application of the Levenshtein similarity index to this particular issue.

It may seem a great deal of cumbersome work to carry out a large number of pairwise tests when dealing with more than two sets of strings as in

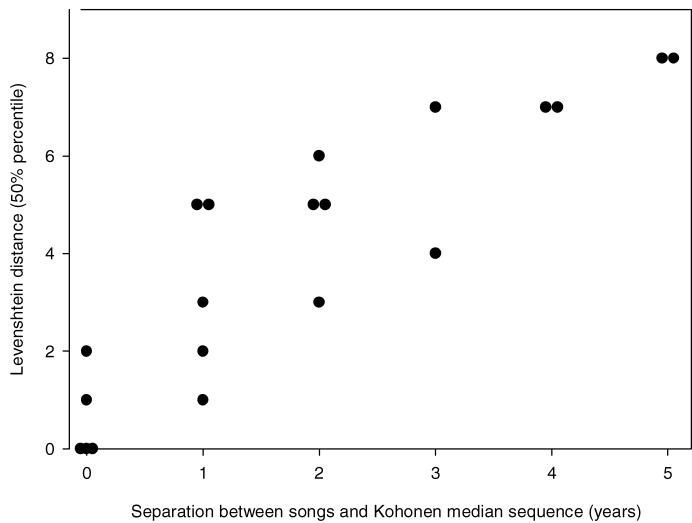


Figure 3. Levenshtein distances (50% percentiles) resulting from all 20 possible combinations of years and Kohonen median sequences plotted as a function of separation in years between songs and Kohonen median.

Table 4. Results of the multiple comparisons of yearly Kohonen median songs and songs from individual years. Shown are Dunn-Sidak corrected *p*-values from the unpaired Mann-Whitney/Wilcoxon tests.

Year	No. of songs	Kohonen median	Year of comparison Kohonen median (<i>p</i> -values)				
			1993	1994	1995	1996	1998
1993	3	ACDEFGEGF	–	1	0.76	0.76	0.76
1994	55	ABCDEFGF	<0.0001	–	<0.0001	<0.0001	<0.0001
1995	22	AA ₂ BCGFH	<0.0001	<0.0001	–	1	<0.0001
1996	36	AA ₂ BCFH	<0.0001	<0.0001	0.0078	–	<0.0001
1998	14	AA ₂ BC	<0.0001	<0.0001	<0.0001	<0.0001	–

the above example. In fact, for *n* sets, we need to complete *n*²-*n* tests, to test all possible pairs. Not only is this a great deal of work, but also the power of the tests will weaken as we increase the number of pairs due to the Bonferroni/Dunn-Sidak correction of probabilities. This correction is conservative in the sense that it maintains a constant probability of committing a type I error (level of significance), but does so at the expense of the power (probability of committing a type II error). The more pairs we compare, the more likely it is that we will miss true differences among the pairwise com-

parisons. From a first inspection it seems more efficient to conduct a one-way ANOVA, or a non-parametric equivalent, followed by an appropriate post-hoc pair-wise comparison, such as Tukey-Kramers test. This approach is not advisable however. A one-way ANOVA test on the sets (more specifically, the sets of associated distances to the Kohonen medians) will not only test for differences between the sets of our interest, but also for other differences not relevant in this context. In fact, most of the comparisons will be of the latter type. These comparisons are not testing differences among years but rather magnitude of differences, such as testing whether the change from year A to year B is different in magnitude from the change from year C to year D. For N groups of songs, there will be N^2 possible sets of distances to Kohonen median songs, which will translate into $N^2(N^2 - 1)$ pairwise comparisons in the subsequent Tukey-Kramer test. In the Humpback whale example, this would translate into 1260 pairwise comparisons, of which 1230 would be irrelevant for the simple question of differences among years. Thus the method presented here seems to have a fundamental limitation to the number of groups it is practically possible to include in the analysis and still maintain reasonable power in the tests. Further work on overcoming this constraint is encouraged, possibly along the lines suggested by Sokal & Rohlf (1995), p. 229 on performing only a limited number of pre-defined pairwise comparisons following a one-way ANOVA.

Hopefully it has been demonstrated in the above that the Levenshtein, or string edit distance can provide a powerful tool for making quantitative statements on differences among groups of songs, provided they can be accurately described by sequences of well-defined units. A first approach to statistical testing of the central hypothesis that two groups of songs belong to the same population of songs has been provided, made possible through the use of median song sequences. The method has already proved its merits on songs from Humpback whales (Helweg et al., 1998; Eriksen et al., 2005) but application to other areas, such as bird song is encouraged, as it provides simple, robust and objective measure of song similarity.

Acknowledgements

We are most grateful to His Majesty King Taufa'ahau Tupou IV, Kingdom of Tonga, for his permission to the South Pacific Humpback Whale Consortium (www.sbs.auckland.ac.nz/research/ecolevol/baker/baker.htm) to record songs of his humpback whales. David A. Helweg is thanked for providing access to the recordings and for constructive comments to the

manuscript. Comments from Frank Riget and an anonymous referee are also appreciated. JT was funded by the Danish National Research Foundation, through the Centre for Sound Communication.

References

- Chabot, D. (1988). A quantitative technique to compare and classify humpback whale (*Megaptera novaeanglia*) sounds. — *Ethology* 77: 89-102.
- Eriksen, N., Tougaard, J., Miller, L.A. & Helweg, D.A. (2005). Cultural change in the songs of humpback whales (*Megaptera novaeangliae*) from Tonga. — *Behaviour* 142: 305-325.
- Helversen, D. & Helversen, O. (1998). Acoustic pattern recognition in a grasshopper: processing in the time or frequency domain? — *Biol. Cybern.* 79: 467-476.
- Helweg, D.A., Cato, D.H., Jenkins, P.F., Garrigue, C. & McCauley, R.D. (1998). Geographic variation in south pacific humpback whale song. — *Behaviour* 135: 1-27.
- Holy, T.E. & Guo, Z. (2005). Ultrasonic songs of male mice. — *PLoS Biology* 3 (12): 1-10.
- Jenkins, P.F. (1977). Cultural transmission of song patterns and dialect development in a free-living bird population. — *Anim. Behav.* 25: 50-78.
- Kohonen, T. (1985). Median strings. — *Pattern Recogn. Lett.* 3: 309-313.
- Kohonen, T. (1988). Self-organization and associative memory 2nd. ed. — Springer Verlag, Berlin.
- Payne, R.S. & McVay, S. (1971). Songs of humpback whales. — *Science* 173: 585-597.
- Sankoff, D. & Kruskal, J.B. (1983). Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. — Addison-Wesley, Reading, Mass.
- Schul, J. (1998). Song recognition by temporal cues in a group of closely related bushcricket species (genus *Tettigonia*). — *J. Comp. Physiol. A* 183: 401-410.
- Sokal, R.R. & Rohlf, F.J. (1995). Biometry: the principles and practice of statistics in biological research 3rd. ed. — W.H. Freeman, New York.
- Thorpe, W.H. (1954). The process of song-learning in the chaffinch as studied by means of the sound spectrograph. — *Nature* 173: 465-469.
-